INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2025, VOL. 71, NO. 3, PP. 1-8
Manuscript received May 21, 2025; revised July 2025. doi: 10.24425/ijet.2025.153609

A comparative study of deep End-to-End Automatic Speech Recognition models for doctor-patient conversations in Polish in a real-life acoustic environment

Karolina Pondel-Sycz, Piotr Bilski, Piotr Bobiński, Leszek Morzyński, Marcin Lewandowski, Emil Kozłowski, Grzegorz Szczepański, Maciej Jasiński, Grzegorz Makarewicz, Agnieszka Paula Pietrzak, Andrzej Buchowicz, Paweł Mazurek, Adrian Bilski, Jacek Olejnik, and Iwona Olejnik

***Abstract*—The following paper presents research on the Automatic Speech Recognition (ASR) methods for the construction of a system to automatically transcribe the medical interview in Polish language during a visit in the clinic. Performance of four ASR models based on Deep Neural Networks (DNN) was evaluated. The applied structures included XLSR-53 large, Quartznet15x5, FastConformer Hybrid Transducer-CTC and Whisper large. The study was conducted on a self-developed speech dataset. Models were evaluated using Word Error Rate (WER), Character Error Rate (CER), Match Error Rate (MER), Word Accuracy (WAcc), Word Information Preserved (WIP), Word Information Lost (WIL), Levenshtein distance, Jaro - Winkler similarity and Jaccard index. The results show that the Whisper model outperformed other tested solutions in the vast majority of the conducted tests. Whisper achieved a WER = 20.84%, where XLSR-53 WER = 67.96%, Quartznet15x5 WER = 76.25%, FastConformer WER = 46.30%. These results show that Whisper needs further adaptation for medical conversations, as current volume of transcription errors is not practically acceptable (too many mistakes in the description of the patient's health description).**

***Keywords*—Automatic Speech Recognition; Transformers; Encoder-Decoder; Deep Neural Network**

I. Introduction

ARTIFICIAL Intelligence (AI) has gained significant interest in recent years, due to its increasing abilities to automate human operations in multiple applications. One of the most prominent is the speech-to-text processing, which allows for accelerating the process of creating text documents and filling the forms. This is important, for instance, in the daily procedures of the medical clinics, which face the problem of the overwhelming paperwork. Transition to the electronic documents allows for proposing the flexible framework for the document flow. One of its possible applications is the automated form-filling, based on the Real-Time speech recording. The typical implementation of the system is two-staged: speech-to-text (aimed at extracting all words) and Natural Language Processing (NLP) for extracting keywords, combining them through context and uploading them to the form.

Two main approaches to Automatic Speech Recognition (ASR) are conventional (using separate models for acoustics, linguistics and lexicon) and End-to-End (E2E), which uses an integrated model based on Deep Neural Networks (DNNs). The latter mainly uses Recurrent (RNN), Convolutional (CNN) and Transformer networks [1].

The aim of this research was to find the most effective Automatic Speech Recognition (ASR) methods that can be used to build a system for transcription of the doctor-patient conversation in Polish language. Examples of ASR methods application in Polish medical terminology are present in literature. They are mainly commercial systems such as Google ASR, Microsoft ASR and Techno ASR [2]. Consideration of Whisper model application to medical terminology was presented in [3] (without demonstration of implementation results). Analysis of these solutions using a corpus of single words was presented in [4], [5]. The more traditional Hidden Markov Model was described in [6]. The following paper examines selected ASR DNN models with the intention to train them on the specifically prepared dataset of doctor-patient conversations in a medical interview scenario.

In such a scenario, ASR models face challenges like ambient sounds, noise, reverberation and overlapping conversations (cocktail party scenario) [7]. Utterances of words and phonemes for both parties may overlap. Also, another person - such as nurse entering the office during a medical visit - can join the conversation. The system is intended to be used as an

This work was supported by the National Centre for Research and Development under Project NFOSTRATEG-IV/0042/2022.

K. Pondel-Sycz, P. Bilski, P. Bobiński, M. Lewandowski, M. Jasiński G. Makarewicz A. P. Pietrzak, A. Buchowicz and P. Mazurek are with Faculty of Electronics and Information Technology, Warsaw University of Technology, Poland (e-mail: {karolina.sycz, piotr.bilski, piotr.bobinski, marcin. lewandowski, maciej.jasinski, grzegorz.makarewicz, agnieszka.pietrzak, andrzej.buchowicz, pawel.mazurek} @wp.edu.pl).

L. Morzyński, E. Kozłowski, G. Szczepański are with Central Institute For Labour Protection-National Research Institute, Poland (email: {lmorzyns, emkoz, grszc} @ciop.pl).

A. Bilski is with Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences, Poland (email: adrian_bilski@sggw.edu.pl).

J. Olejnik and I.Olejnik are with JAS Technologie Sp. z o.o, Poland (email: j.olejnik@jastechnologie.pl, iwonaolejnik10@wp.pl).

application installed on a computer available in the clinic, so the recording quality will depend on the available hardware (recordings from a considerable distance may be of low quality).

The study investigated how the content of medical terminology (names of drugs, diseases and symptoms) and the quality of recordings (acoustic conditions, quality of audio tools used) affect the performance of E2E ASR-based Polish speech recognition. Four DNN models have been employed for this task: XLSR-53 large Polish [8], STT Pl Quartznet15x5 [9], STT Pl FastConformer Hybrid Transducer-CTC Large P&C [10] and Whisper large [11].

The models were tested by comparing the transcription obtained from their output (hypothesis) with the actual text contained in the recording (reference) on the basis of indicators: Word Error Rate, Character Error Rate, Match Error Rate, Word Information Preserved, Levenshtein distance, Jaro - Winkler similarity and Jaccard index. Chapter II describes the planned application of the ASR model in a doctor-patient conversation recognition system, Chapter III details the architectures of the models tested, the evaluation metrics used and the statistical methods. Chapter IV contains a description of the authors' dataset used to test the selected models, Chapter V reports the experimental process and the results obtained, as well as the observed limitations. Chapter VI contains a summary of the experiments conducted and conclusions.

II. POLISH ASR FOR AUTOMATED ANALYSIS OF DOCTOR-PATIENT CONVERSATIONS

The E2E ASR systems used in the presented scenario belong to one of two groups: universal models also considering Polish, or the specific, language-oriented solutions. In most cases the former are used, as in the system constructed for the presented project (Fig. 1).

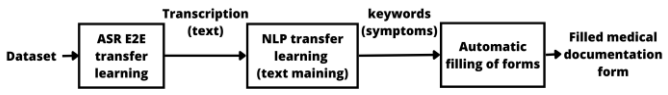

Fig. 1. Designed system for automatic supplementing medical records based on doctor-patient interview during medical examination.

Polish language, which is the subject of the analysis in the developed system, has specific requirements. It is an inflectional, characterized by a rich morphology in which words (nouns, pronouns, adjectives and verbs) conjugate according to grammatical context (by case, gender, person, or plurality). This results in dictionaries containing hundreds of thousands of words. It also affects the acoustic model of the language - subtle phonetic differences between suffixes lead to similarly-sounding words [12].

Sentence construction in Polish is relatively unrestricted and does not rely heavily on the Subject-Verb-Object (SVO) formation. The meaning is often determined by the complex inflection, and the function of a word depends on its form rather than the position in the sentence [13], [14], [15], [16].

In terms of spelling, Polish contains consonant combinations that are pronounced as single sounds, such as “cz,” “sz,” “dz” and “dż”, and sounds with multiple spellings, such as “ż/rz,” “h/ch” and “u/ó.” Moreover, the pronunciation of “ć/ci” varies depending on the spelling [17], [18], [19].

III. APPLIED METHODS

This section includes a description of the E2E DNN ARS models used in the study, adapted for Speech-To-Text (STT) recognition, and methods for evaluating them in recognizing content of the medical interview.

A. Architectures of used E2E DNN ASR models

For the study, Open-Source E2E ASR models were selected in the hope of gaining new knowledge from the additional data set through the transfer learning. Among the selected structures, three run in the Polish language mode (further referred to as *polish version* models) without a language detection stage STT Pl Quartznet15x5 (referred as Quartznet) and STT Pl FastConformer Hybrid Transducer-CTC Large P&C (referred as FastConformer) and multilingual models XLSR-53 large Polish (referred as Wav2Vec) and Whisper. All selected models represent the Encoder-Decoder (ED) architecture, common in Sequence-to-Sequence (Seq2Seq) tasks. The encoder's task is to convert the input sequence into vectors of the appropriate size, while the decoder is supposed to return the most probable labels based on the feature vectors taken from the encoder.

1) STT Pl Quartznet15x5

The NVIDIA Quartznet [20] builds upon the Jasper model [21] incorporating CNN layers trained with a CTC (Connectionist Temporal Classification) loss function. It introduces the BxR architecture, featuring B blocks, each containing R convolutional sub-blocks. Quartznet utilizes spectrograms as input speech features.

A key enhancement in this architecture is the substitution of 1D convolutions, as used in Jasper, with 1D Time-Channel Separable Convolutions (1DTCSC), where "time" pertains to one-dimensional data. The convolution filter moves along the time axis and the operation produces temporal and channel components. The former employ distinct convolution filters for each time step, facilitating the analysis of temporal data from diverse viewpoints, enabling the detection of patterns and relationships. Channel operations splice the resultant data, extracting features across multiple channels. A variant of Quartznet15x5 was used, consisting of 15 blocks, each composed of the same base modules repeated 5 times (which included four layers: depth convolution, pointwise convolution, normalization and ReLU). This architecture is accessible via the NVIDIA NeMo toolkit [22].

In the selected language version of the model, the Encoder is adopted from the English version of Quartznet, while the decoder is modified to generate characters from the Polish alphabet. The adaptation was done using the Polish fragment of the Mozilla Common Voice (MCV) dataset [23], [24].

2) STT Pl FastConformer Hybrid Transducer-CTC Large P&C

This model (also sourced from the NVIDIA NeMo Toolkit) was adapted for Polish using three datasets: MCV, Multilingual LibriSpeech (MLS) [25], and VoxPopuli (VP) [26]. The model employs a FastConformer architecture [27], [28] with a Transducer (extension of CTC with joint modeling of input-output and output-output relationships) [29] and CTC decoder loss. FastConformer employs an encoder modification, in which

the sampling rate is increased from 10 ms to 80 ms by using 3 layers of convolutional depth subsampling. The additional 2x reduction in encoder output length provides computational memory savings in the decoder.

The FastConformer version selected for the study uses a hybrid decoder, combining a Recurrent Neural Network Transducer (RNN-T) and a CTC. It employs the Google SentencePiece Unigram tokenizer [30] and transcribes text in both uppercase and lowercase letters, including spaces, periods, commas, question marks, and several other characters.

3) XLSR-53 large Polish

The model derives from a platform developed by Facebook AI (Meta) for self-learning speech representations based on CNNs and Transformer networks. The raw speech waveform is fed to the input of the CNN encoder, which produces its hidden representations at the output. These are processed by the Transformer encoder, which output is processed by the quantization module to represent targets for self-supervised learning. The model builds context representations on continuous speech [31]. In [32], the use of the Wav2Vec 2.0 framework for unsupervised learning of the ASR XLSR-53 multilingual model is described, covering 53 languages (including Polish). Tuning the model to new languages is performed by training it on additional data with CTC loss function [33]. In the presented case, MCV, MLS and Babel [34] were used.

4) Whisper

The Whisper is OpenAI's open-source, multilingual model, built on the Transformer Encoder-Decoder (ED) [35] architecture with two additional CNN layers at the top of the encoder's structure (referred to as Speech-Transformer). It supports 57 languages, including Polish [11].

The model comes in five versions: tiny, base, small, medium, and large. It allows for direct mapping between utterances and their transcriptions by predicting the raw text, eliminating the need for significant standardization or preprocessing.

The used model has been trained on a large set of audio data and transcriptions from the Internet. It exhibits diversity in terms of audio quality, which helps becoming robust to changes in speech signals [36].

B. Evaluation

The capability of E2E DNN ASR models is usually assessed by two metrics: Word Error Rate (WER) and Character Error Rate (CER). This section introduces other metrics that can be used to evaluate the system. All of them evaluate the quality of the model's output transcription (hypothesis) by comparing it to a reference text.

1) Word Error Rate (WER)

It is a standard approach to evaluating the performance of an ASR model. It evaluates a ratio of the sum of errors in the model hypothesis (substitutions (S), insertions (I) and deletions (D)) to the overall number of words in the reference N (eq. 1).

$$WER = \left(\frac{S+I+D}{N}\right) \cdot 100\% \qquad (1)$$

The greater WER indicates the presence of increased errors in the model's hypothesis. It is desirable to get the smallest possible value [37]. The WER is provided in practice as an average across the test set.

2) Character Error Rate (CER)

It is an error rate similar to WER, but operating at the level of individual characters instead of whole words. The CER equation is the same as eq. 1, but considers characters.

3) Match Error Rate (MER)

MER is the probability that the word match between the hypothesis and the reference is incorrect. Like WER, taking into account word-level S, D, and I errors [38]:

$$MER = \left(\frac{S+I+D}{N+I}\right) \cdot 100\% \qquad (2)$$

4) Word Information Preserved (WIP)

The measure indicates the percentage of words correctly predicted between the reference and the hypothesis. This is an accuracy measure, so a higher value indicates better performance of the ASR model [38]:

$$WIP = \frac{H}{N_1} \cdot \frac{H}{N_2} \qquad (3)$$

where H is the number of correctly recognized words, N_1- sum of words in reference and N_2- sum of words in hypothesis.

5) Levenshtein distance

This measure [39] operates on the characters. It represents the minimum number of basic operations (I, D or S) required to transform one string into another. The distance $Lev.dist_{u,v}$ between the strings *u* and *v* is [40]:

$$Lev.dist_{u,v}(i,j) = \begin{cases} max(i,j) & I \\ min\begin{cases} d(i-1,j)+1 \\ d(i,j-1)+1 \\ d(i-1,j-1)+\mathbf{1}_{(u_i \neq u_j)} \end{cases} & II \end{cases} \qquad (4)$$

where: I - *if min(i, j) = 0*; II - otherwise; $Lev.dist_{u,v}(i,j)$ - Levenshtein distance between the first *i* characters of the string *u* and the first *j* characters of string *v*; $\mathbf{1}_{(u_i \neq u_j)}$ - function, which is 1 if the characters u_i and v_j are different, and 0 if they are the same.

6) Jaro – Winkler similarity (JW)

It is a modification of the Jaro similarity [41], [42] considering the length of the common prefix. It is used to calculate the similarity *J(u,v)* between two strings *u* and *v* of lengths Len_u and Len_v:

$$J(u,v) = \begin{cases} 0 & I \\ \frac{1}{3}\left(\frac{m}{|u|} + \frac{m}{|v|} + \frac{m-t}{m}\right) & II \end{cases} \qquad (5)$$

where I - if m = 0; II - for m ≠ 0; m - number of the same characters in both strings; t - number of rearranged characters (transpositions).

The Jaro-Winkler similarity introduced a prefix scale factor *p*, which gives more precise results when strings share a common prefix up to a specified maximum length *L* [43]:

$$JW(u,v) = J(u,v) + p \cdot L \cdot (1 - J(u,v)) \qquad (6)$$

The Jaro-Winkler similarity between two words simply shows how close they are.

7) Jaccard index

This emasure assesses the distance between two sets U and V [44]:

$$Jaccard(U,V) = \frac{|U \cap V|}{|U \cup V|} \quad (7)$$

In the context of hypothesis and reference, Jaccard index can be used to assess how similar they are to each other based on how many shared words exist relative to all words.

C. Statistics methods

The mean, standard deviation, box plots and ANOVA test were used to analyze the results obtained from the tests.

Box plots allow for visualization of data distribution, concentration, symmetry and detection of outliers. It enables visual comparison of the obtained metrics values between models. Median and quantiles are used to assess the central tendency of the data and its dispersion.

The ANOVA test (Analysis of Variance) allows for comparison of means (of two or more groups), identification of factors with significant influence on the dependent variable, multivariate analysis (influence of multiple factors simultaneously). In this study, the ANOVA test was used to compare the mean values of the metrics obtained for the different types of microphones used in the dataset recordings.

IV. DATASET

The self-developed dataset was used in the test. It consists of medical interviews recorded in a laboratory room similar to a typical doctor's office, based on a reenactment of a doctor-patient conversation. The dataset is characterized by realistic acoustic conditions of the medical office (presence of ambient sounds, noise, reverberation), contains overlapping spontaneous speech (with mistakes, repetitions, interruptions, breaths, etc.), medical terminology (names of drugs, diseases, symptoms). Speakers are both female and male, speaking Polish. Conversations are held in a configuration: female doctor - female or male patient.

The dataset contains 82 recordings along with transcriptions with time-stamping (start and end) of each speaker's turn. They were acquired using microphones of varying quality (each interview was recorded with two or three different microphones), In total, 5 different microphones were used, marked as Z02HV, Z05BO, Z06GR (omnidirectional, USB-based, cheap), Z10BK (precise free field measurements). Length of recordings varies from 1.36 to 6.34 mins, with an average of 2.97 mins, 4.05h total. Originally the recordings were acquired as two-channel, but due to the requirements of the NeMo Toolkit models, they were converted to the mono format with sample rate 16 kHz.

Since the recordings include patient-doctor dialogues and transcriptions with speaker turns and time stamps, they can be divided into about 2,060 utterances.

V. EXPERIMENT

The experiment consisted of testing off-the-shelf models, described in Section III, unchanged as they were provided by the authors. In the study, we did not perform additional training of the models. Our goal is to preliminarily test the performance of the selected methods in the task of ASR of Polish speech with medical terminology and to identify the most promising solution for further adaptation to the planned system (Figure 1). To achieve this, all recordings in the dataset were fed to the input of each model, without additional preprocessing. The output was a hypothesis in the textual form. The reference text and the hypothesis were normalized: transforming characters to uppercase and removing all special characters. The reference texts and hypotheses produced in this manner were used to calculate the metrics listed in Section III-B. Average values of metrics were calculated along with the standard deviation, box plots for each metric were generated, and an ANOVA test of the effect of microphone quality on the result of recognition (hypothesis) was conducted.

The average values of the metrics along with the standard deviation are in Table I. box plots refer respectively to: WER (Fig. 2), CER (Fig 3), MER (Fig. 4), WIP (Fig. 5), Levenshtein distance (Fig. 6), Jaro-Winkler similarity (Fig. 7) and Jaccard index (Fig. 8) and the average metrics values for each of microphone used (Tables II, III, IV, V).

Error and inaccuracy metrics such as WER, CER, MER, and Levenshtein distance should be interpreted as: the smaller the value of the metrics, the smaller the number of errors or the accuracy of mapping the reference text by the hypothesis. Efficiency and correctness metrics such as WIP, JW sim. and Jaccard should be interpreted reciprocally - a higher indicator value means a better representation of the reference by the hypothesis.

Whisper performed best for all metrics, except JW. sim., for which the result is close to the Wav2Vec. Analyzing its box plot, it can be seen that Whisper's results are less dispersed, so the model makes mistakes, but on a smaller scale. Its counterparts are less predictable in this matter. Analyzing the box plot and average Jaccard index for Whisper, it can be seen that in most cases the reference and hypothesis strings are close to each other. The advantage of the Whisper model is also highlighted by the Levenshtein distances obtained, which are lower than for the other models. This indicates that fewer edits to the hypothesis are required to obtain the same text as in the reference. From all box plots it can be seen that for single samples (outliers) Whisper made mistakes comparable to the other models, but overall it got better results and the values of the metrics were less dispersed around the median. The minimum WER value for Whisper was 6.97%, which indicates that it is able to achieve an efficiency close to that reported by its developers.

Wav2Vec, while scoring higher WER and MER, at the same time its CER and Levinstein Distance are smaller compared to FastConformer and Quartznet models. This may indicate that the model makes typos in a larger number of words, which affects the large WER value (if we consider the words as a whole they are wrong), but there are more accurate mappings in the hypotheses at the level of single characters and close strings. The FastConformer and Quartznet models achieved similar CER, but in terms of word hits and correct matches (WER and MER), the FastConformer model was significantly better. Based on all metrics, Quartznet performed the weakest, as highlighted by a box plot analysis of MER and WIP, from which it can be seen that the model incorrectly predicts at a constant and high level.

From an analysis of the average values of the metrics by the microphone used in the recording, it can be seen that the best

recognition results were obtained for the recordings made with the professional measurement microphone, and the averages obtained for this microphone are close to or better than the average values of the metrics calculated for all the results obtained (Table I). Thus, the effect of microphone quality on model performance is noticeable, but using the WER as an example, the benefit of using it, ranges from about 1.5-6%. The greatest benefit of using this microphone can be seen in the WER of the Wav2Vec, and the least impact is observed for Whisper (a decrease in WER of less than 1.5%).

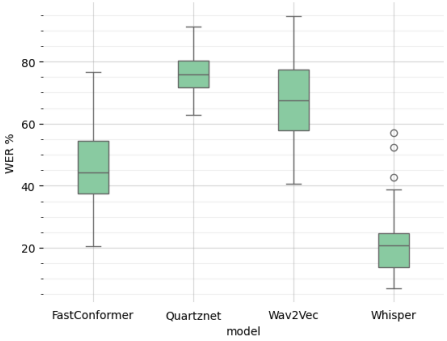

Fig. 2. WER [%] box plot.

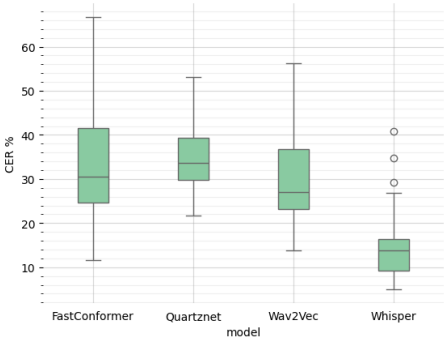

Fig. 3. CER [%] box plot.

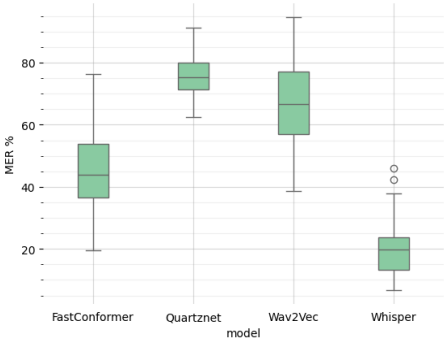

Fig. 4. MER [%] box plot.

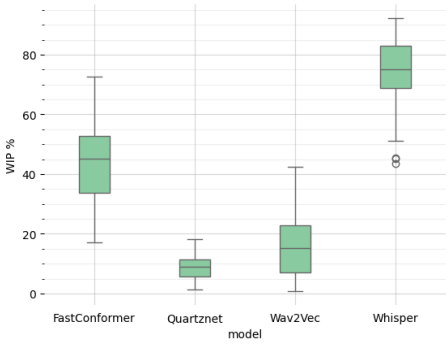

Fig. 5. WIP [%] box plot.

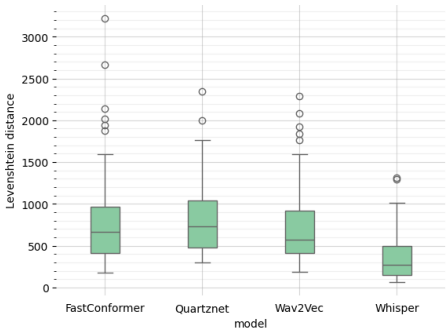

Fig. 6. Levenshtein distance box plot.

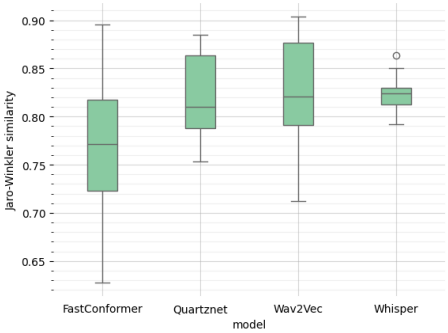

Fig. 7. Jaro-Winkler similarity box plot.

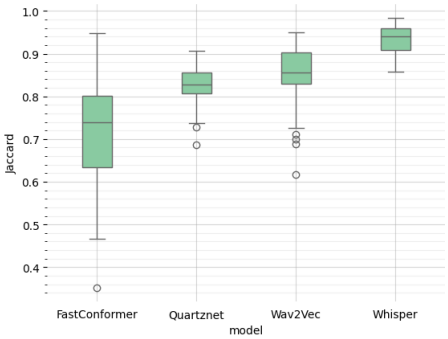

Fig. 8. Jaccard index box plot.

Before performing the ANOVA test, it was verified that provided data represent the universe (space containing all possible words) with a normal distribution. Also, standard deviations of the groups in the population are equal (homoscedasticity). Not all obtained metric values meet these assumption. For FastConformer, the assumptions of the ANOVA test are met for metrics WER, MER ($p \leq 0.02$), WIP, ($p \leq 0.01$), JW sim. ($p \leq 0.23$) and Jaccard ($p \leq 0.20$). For Quartznet, these assumptions are fulfilled only for the metrics WER, MER ($p \leq 0.01$). For the Wav2Vec, these assumptions are satisfied for all metrics (except Lev. dist., JW sim. and Jaccard) and for these metrics $p \leq 0.01$. For Whisper, these assumptions are met only for the metrics JW sim. ($p \leq 0.76$) and Jaccard ($p \leq 0.84$).

TABLE I
AVERAGE VALUES (WITH A STANDARD DEVIATION) OF METRICS FOR TESTED MODELS

Metrics	FastConformer	Quartznet	Wav2Vec	Whisper
WER %	46.30 (±12.33)	76.25 (±6.88)	67.96 (±12.54)	20.84 (±9.53)
CER %	33.18 (±12.29)	34.89 (±7.47)	29.87 (±9.55)	13.89 (±6.59)
MER %	45.66 (±12.36)	75.88 (±7.00)	76.08 (±12.91)	9.68 (±8.23)
WIP %	43.67 (±12.50)	8.77 (±4.03)	15.74 (±9.81)	73.99 (±10.96)
Lev. Dist.	808.96 (±559.80)	830.10 (±433.27)	730.68 (±463.49)	356.40 (±271.86)
JW sim.	0.77 (±0.06)	0.82 (±0.04)	0.83 (±0.05)	0.82 (±0.01)
Jaccard	0.72 (±0.13)	0.83 (±0.04)	0.85 (±0.06)	0.93 (±0.03)

TABLE II
AVERAGE METRICS VALUES (WITH A STANDARD DEVIATION) FOR FASTCONFORMER MODEL PER MICROPHONE TYPE. TYPE

Metrics	Z02HV	Z03AF	Z05BO	Z06GR	Z10BK
WER %	53.34 (±12.48)	56.33 (±9.15)	50.12 (±17.20)	45.05 (±11.03)	42.22 (±10.32)
CER %	41.55 (±12.01)	41.50 (±9.15)	35.30 (±16.80)	32.33 (±11.63)	29.48 (±10.39)
MER %	52.59 (±12.34)	55.64 (±9.29)	49.75 (±17.43)	44.39 (±10.98)	41.54 (±10.30)
WIP %	37.33 (±12.27)	31.22 (±9.05)	38.74 (±17.32)	45.39 (±10.40)	48.13 (±10.58)
Lev. Dist.	891.33 (±642.78)	873.71 (±397.36)	1060.36 (±867.83)	785.31 (±542.88)	704.59 (±442.79)
JW sim.	0.73 (±0.05)	0.75 (±0.05)	0.78 (±0.08)	0.77 (±0.06)	0.79 (±0.06)
Jaccard	0.64 (±0.13)	0.66 (±0.11)	0.70 (±0.18)	0.73 (±0.13)	0.76 (±0.11)

TABLE III
AVERAGE METRICS VALUES (WITH A STANDARD DEVIATION) FOR QUARTZNET MODEL PER MICROPHONE TYPE. TYPE

Metrics	Z02HV	Z03AF	Z05BO	Z06GR	Z10BK
WER %	81.54 (±4.19)	85.72 (±5.45)	79.25 (±8.30)	74.81 (±5.49)	73.18 (±5.34)
CER %	42.09 (±3.62)	46.69 (±5.57)	38.08 (±9.81)	32.77 (±5.13)	31.45 (±5.00)
MER %	80.93 (±4.14)	85.66 (±5.49)	79.04 (±8.45)	74.41 (±5.55)	72.74 (±5.49)
WIP %	5.59 (±2.51)	3.62 (±2.32)	6.98 (±4.29)	9.45 (±3.35)	10.66 (±3.54)
Lev. Dist.	843.83 (±461.60)	974.71 (±412.28)	1100.36 (±626.10)	785.31 (±542.88)	745.14 (±368.92)
JW sim.	0.69 (±0.03)	0.64 (±0.05)	0.71 (±0.08)	0.76 (±0.04)	0.77 (±0.04)
Jaccard	0.81 (±0.01)	0.77 (±0.06)	0.81 (±0.06)	0.84 (±0.03)	0.84 (±0.03)

TABLE IV
AVERAGE METRICS VALUES (WITH A STANDARD DEVIATION) FOR WAV2VEC MODEL PER MICROPHONE TYPE. TYPE

Metrics	Z02HV	Z03AF	Z05BO	Z06GR	Z10BK
WER %	84.07 (±9.91)	79.91 (±7.73)	74.28 (±11.57)	65.41 (±10.57)	61.89 (±10.64)
CER %	45.43 (±9.39)	40.00 (±7.13)	33.45 (±9.87)	27.60 (±6.80)	25.11 (±6.70)
MER %	83.40 (±10.43)	79.28 (±8.47)	73.70 (±11.93)	64.45 (±10.57)	60.88 (±10.64)
WIP %	5.40 (±5.26)	7.20 (±4.87)	10.65 (±7.91)	17.20 (±8.75)	20.39 (±9.66)
Lev. Dist.	942.67 (±644.01)	846.86 (±397.91)	978.09 (±604.17)	676.69 (±418.47)	618.93 (±393.66)
JW sim.	0.74 (±0.06)	0.79 (±0.05)	0.83 (±0.05)	0.87 (±0.04)	0.88 (±0.04)
Jaccard	0.81 (±0.10)	0.77 (±0.07)	0.81 (±0.06)	0.84 (±0.04)	0.84 (±0.04)

TABLE V
AVERAGE METRICS VALUES (WITH A STANDARD DEVIATION) FOR WHISPER MODEL PER MICROPHONE TYPE. TYPE

Metrics	Z02HV	Z03AF	Z05BO	Z06GR	Z10BK
WER %	23.87 (±14.60)	26.56 (±8.31)	23.03 (±7.79)	19.40 (±10.13)	19.45 (±8.33)
CER %	15.57 (±9.68)	17.16 (±6.04)	15.48 (±5.51)	12.97 (±7.16)	13.06 (±5.77)
MER %	21.65 (±10.90)	25.48 (±8.15)	21.84 (±7.23)	18.23 (±8.36)	18.51 (±7.57)
WIP %	70.36 (±13.36)	63.88 (±12.31)	71.16 (±10.58)	76.51 (±10.08)	75.74 (±10.00)
Lev. Dist.	375.00 (±453.58)	364.29 (±205.44)	455.82 (±279.29)	337.55 (±280.61)	331.79 (±237.60)
JW sim.	0.82 (±0.01)	0.82 (±0.01)	0.82 (±0.01)	0.82 (±0.01)	0.82 (±0.01)
Jaccard	0.93 (±0.04)	0.93 (±0.04)	0.93 (±0.03)	0.94 (±0.03)	0.94 (±0.03)

The main limitation of the study is that not all the obtained values of metrics, divided according to the microphone used, meet the assumptions of the ANOVA test. This might be due to too small sample size in the population. Therefore it was necessary to exclude ANOVA test results from some tests.

Selection of open-source E2E DNN ASR models capable of recognizing Polish is also limited. To the best of our knowledge, in addition to the presented models, there is a multilingual model in the ESPnet toolkit that is adapted to the Polish language. However, preliminary tests show that it achieves a WER of 90%. The tests also show limitations of these models in the task of recognizing Polish speech in realistic acoustic conditions, a spontaneous conversation with speakers overlapping.

For our dataset, WER of all tested models differs significantly from the values provided by the developers. Even for the best one (Whisper) in our tests WER is about 15% higher than the value provided in the specification (6%). It is therefore not possible to use any models for medical applications and requires fine-tuning them with new datasets.

It is then important to consider limitation of the text normalization performed, which only included changing all

letters to uppercase and removing punctuation marks. The process did not consider converting numbers into words, which may have negatively affected the obtained WER results. The reason for this is that in Polish, numbers are conjugated by genders and cases, so one numerical notation has multiple forms of word notations.

CONCLUSION

The study tested 4 deep neural models with different architectures and training set (Quartznet, FastConformer, XLSR-53 large, Whisper-larg) capable of recognising speech in Polish, in a doctor-patient conversation recognition task, based on a self-developed data set. Models were assessed on the basis of measures of WER, CER, MER, WIP, Levenshtein distance, Jaro - Winkler similarity and Jaccard index. Mean values, standard deviations, aggregation around the median and the dependence of the metrics on the mycrophone used in the recordings were measured. Whisper performed best for all metrics, except JW. sim., for which the result is close to the Wav2Vec. The values of all Whisper metrics are less dispersed around the median, meaning that the model makes errors on a narrower scale. It also showed small differences in the performance quality of this model, depending on the class of microphone used for the recording (a decrease in WER of less than 1.5% for a high-quality microphone), indicating the resilience of this model to the quality of the recording. However, for a data set unknown to the model, its recognition quality decreases significantly (compared to the Mozilla Common Voice9 database, for which WER=9.0% [32]). It is therefore not possible to use any of the models for the proposed system. All models require tuning using new datasets.

ACKNOWLEDGEMENTS

This work was supported by the Polish National Center for Research and Development (NCBiR) under the INFOSTRATEG IV project (Intelligent speech processing system for medical professionals).

REFERENCES

[1] J. Li et al., “Recent advances in end-to-end automatic speech recognition,” arXiv (Cornell University), 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2111.01690

[2] K. Kuligowska, M. Stanusch, and M. Koniew, “Challenges of automatic speech recognition for medical interviews-research for polish language,” Procedia Computer Science, vol. 225, pp. 1134–1141, 2023.

[3] A. Czyżewski, “Optimizing medical personnel speech recognition models using speech synthesis and reinforcement learning,” Journal of the Acoustical Society of America, vol. 154, pp. A202–A203, 2023.

M. Zielonka, W. Krasiński, J. Nowak, P. Rośleń, J. Stopiński, M. Żak, F. Górski, and A. Czyżewski, “A survey of automatic speech recognition deep models performance for polish medical terms,” in 2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA). Poznan, Poland: IEEE, 2023, pp. 19 – 24

S. Zaporowski, “The impact of foreign accents on the performance of whisper family models using medical speech in polish,” in Harnessing Opportunities: Reshaping ISD in the post-COVID-19 and Generative AI Era (ISD2024 Proceedings), B. Marcinkowski, A. Przybylek, A. Jarzebowicz, N. Iivari, E. Insfran, M. Lang, H. Linger, and C. Schneider, Eds. Gdańsk, Poland: University of Gdańsk, 2024

B. Hnatkowska and J. Sas, “Application of automatic speech recognition to medical reports spoken in polish,” Journal of Medical Informatics Technologies, vol. 12, 2008

T.-B. Nguyen and A. Waibel, “Convoifilter: A case study of doing cocktail party speech recognition,” arXiv (Cornell University), Aug. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2308.11380

[4] Grosman, Jonatas, “Fine-tuned XLSR-53 large model for speech recognition in Polish,” https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-polish 2021, online; accessed 20 May 2024

[5] NVIDIA Nemo Toolkit, “STT Pl Quartznet15x5,” https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_pl_quartznet15x5, 2023, online; accessed 20 May 2024

[6] NVIDIA, “STT Pl FastConformer Hybrid Transducer-CTC Large P&C,” https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_pl_fastconformer_hybrid_large_pc, 2023, online; accessed 20 May 2024.

[7] OpenAI, “Whisper - general-purpose speech recognition mode,” https://github.com/openai/whisper, 2023, online; accessed 20 May 2024.

[8] E. Rudnicka, M. Maziarz, M. Piasecki, and S. Szpakowicz, “A strategy of mapping polish wordnet onto princeton wordnet,” in Proc. 24th COLING 2012: Posters. Mumbai, India: COLING 2012, 8-15 Dec. 2012, pp. 1039––1048. [Online]. Available: https://aclanthology.org/C12-2101/

[9] A. Pohl and B. Zi´ołko, “Using part of speech n-grams for improving automatic speech recognition of polish,” in Machine Learning and Data Mining in Pattern Recognition. Proc. of 9th Int. Conf., MLDM 2013. New York, NY, USA: Springer Berlin, Heidelberg, 19-25 Jul. 2013, pp. 492–504. [Online]. Available: https://doi.org/10.1007/978-3-642-39712-7 38

[10] G. Rehm and H. Uszkoreit, The Polish language in the digital age. Berlin, Heidelberg: Springer, Jan. 2012. [Online]. Available: https://doi.org/10.1007/978-3-642-30811-6

[11] A. Janicki and D. Wawer, “Automatic speech recognition for polish in a computer game interface,” in 2011 Federated Conf. on Computer Science and Information Systems (FedCSIS). Szczecin, Poland: IEEE, 18-21 Sep. 2011, pp. 711—716. [Online]. Available: https://ieeexplore.ieee.org/document/6078265

[12] K. Marasek, Brocki, D. Korzinek, K. Wołk, and R. Gubrynowicz, “Spoken language translation for polish,” arXiv (Cornell University), Nov. 2015. [Online]. Available: doi.org/10.48550/arXiv.1511.07788

[13] L. Pawlaczyk and P. Bosky, “Skrybot – a system for automatic speech recognition of polish language,” in Man-Machine Interactions. Berlin, Heidelberg: Springer, Jan. 2009, vol. 59, pp. 381–387. [Online]. Available: https://doi.org/10.1007/978-3-642-00563-3_40

[14] J. Nouza, P. Cerva, and R. Safarik, “Cross-lingual adaptation of broadcast transcription system to polish language using public data sources,” in Human Language Technology. Challenges for Computer Science and Linguistics. 7th Language and Technology Conf., LTC 2015. Poznań, Poland: Springer Cham, 27-29 Nov. 2015, pp. 31–41. [Online]. Available: https://doi.org/10.1007/978-3-319-93782-3_3

[15] B. Ziółko, S. Manandhar, R. C. Wilson, M. Ziółko, and J. Galka, “Application of HTK to the Polish language,” in ICALIP 2008 Int. Conf. on Audio, Language and Image Processing. Shanghai, China: IEEE, 7-9 Jul. 2008, pp. 31–41. [Online]. Available: https://doi.org/10.1109/ICALIP.2008.4590266

[16] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, “Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions,” in 2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 4-8 May 2020, pp. 6124–6128. [Online]. Available: https://ieeexplore.ieee.org/xpl/conhome/9040208/proceeding?isnumber=9052899&sortType=vol-only-seq&searchWithin=Quartznet

[17] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, “Jasper: An end-to-end convolutional neural acoustic model,” in INTERSPEECH 2019. Graz, Austria: ISCA Speech, 15–19 Sep. 2019, pp. 71–75. [Online]. Available: https://www.isca-archive.org/interspeech_2019/li19_interspeech.html

[18] NVIDIA NeMo: conversational AI toolkit,” (2023). [Online]. Available: https://github.com/NVIDIA/NeMo

[19] “Common Voice open source, multi-language dataset of voices,” (2023). [Online]. Available: https://commonvoice.mozilla.org/en/datasets

[20] J. Huang, O. Kuchaiev, P. O’Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, and B. Ginsburg, “Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition,” in 2021 IEEE Int. Conf. on Multimedia and Expo (ICME). Shenzhen, China: IEEE, -05-09 Jul. 2021, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICME51207.2021.9428334

[21] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in INTERSPEECH 2020. Shanghai, China: ISCA Speech, 25-29 Oct. 2020, pp. 2757–2761. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-2826

[22] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A Large- Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation," arXiv (Cornell University), Jan. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2101.00390

[23] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu et al., "Conformer: Convolution- augmented transformer for speech recognition," in INTERSPEECH 2020. Shanghai, China: ISCA Speech, 25-29 Oct. 2020, pp. 5036–5040. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-3015

[24] D. Rekesh, S. Kriman, S. Majumdar, V. Noroozi, H. Juang, O. Hrinchuk, A. Kumar, and B. Ginsburg, "Fast conformer with linearly scalable attention for efficient speech recognition," arXiv (Cornell University), May 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.05084

[25] A. Graves, "Sequence transduction with recurrent neural networks," arXiv (Cornell University), Nov. 2012. [Online]. Available: https://doi.org/10.48550/arXiv.1211.3711

[26] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," arXiv (Cornell University), Aug. 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1808.06226

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems 33: Annu. Conf. on Neural Information Processing Systems (NeurIPS 2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. Curran Associates, Inc., 6-12 Dec. 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[28] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in INTERSPEECH 2021. Brno, Czechia: ISCA Speech, 30 Aug. - 3 Sep. 2021, pp. 2426–2430. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-329

[29] A. Graves, S. Fern´andez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proc. of the 23rd Int. Conf. on Machine learning. Pittsburgh Pennsylvania USA: Association for Computing Machinery, New York, United States, Jun. 2006, pp. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[30] P. Roach, S. Arnfield, W. Barry, J. Baltova, M. Boldea, A. Fourcin, W. Gonet, R. Gubrynowicz, E. Hallum, L. Lamel et al., "Babel: An eastern european multi-language database," in Proc. of 4th Int. Conf. on Spoken Language Processing (ICSLP 96), vol. 3, Philadelphia, PA, USA, 3-6 Oct. 1996, pp. 1892–1893. [Online]. Available: https://www.isca-archive.org/icslp_1996/roach96_icslp.html#

[31] S. Toshniwal, A. Kannan, C.-C. Chiu, Y. Wu, T. N. Sainath, and K. Livescu, "A comparison of techniques for language model integration in encoder-decoder speech recognition," in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 369–375. [Online]. Available: https://ieeexplore.ieee.org/document/8639038

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. of the 40th Int. Conf. on Machine Learning, ser. Proc. of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. Honolulu, HI, United States: PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[33] A. Ali and S. Renals, "Word error rate estimation for speech recognition: e-wer," in Proc. of the 56th Annu. Meeting of the Association for Computational Linguistics, vol. 2. ACL, 2018, pp. 20–24. [Online]. Available: https://aclanthology.org/P18-2

[34] A. C. Morris, V. Maier, and P. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition," in Proc. INTERSPEECH 2004. ISCA Speech, Oct. 4-8 2004, pp. 2765–2768. [Online]. Available: https://www.isca-archive.org/interspeech_2004/morris04_interspeech.html

[35] V. I. Levenshtein et al., "Binary codes capable of correcting deletions, insertions, and reversals," Soviet Physics Doklady, vol. 10, no. 8, pp. 707–710, 1966.

[36] A. S. Lhoussain, G. Hicham, and Y. Abdellah, "Adaptating the levenshtein distance to contextual spelling correction," International Journal of Computer Science and Applications, vol. 12, no. 1, pp. 127–133, 2015. [Online]. Available: https://www.researchgate.net/publication/273758433_Adaptating_the_Levenshtein_Distance_to_Contextual_Spelling_Correction

[37] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida," Journal of the American Statistical Association, vol. 84, no. 406, pp. 414–420, 1989.

[38] W. E. Winkler, "The state of record linkage and current research problems," Statistical Research Division, US Bureau of the Census, Wachington, DC, 1999. [Online]. Available: https://www.researchgate.net/publication/2509449 The State of Record Linkage and Current Research Problems

[39] O. Rozinek and J. Mares, "Fast and precise convolutional jaro and jaro-winkler similarity," in 2024 35th Conf. of Open Innovations Association (FRUCT). Tampere, Finland: IEEE, 2024, pp. 604–613.

[40] M. Masson and J. Carson-Berndsen, "Investigating phoneme similarity with artificially accented speech," in Proc. of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology. Association for Computational Linguistics, 2023, pp. 49–57. [Online]. Available: https://aclanthology.org/2023.sigmorphon-1